Federated Composite Optimization

Honglin Yuan [1 2] **Manzil Zaheer** [3] **Sashank Reddi** [3]

Abstract

Federated Learning (FL) is a distributed learning paradigm that scales on-device learning collaboratively and privately. Standard FL algorithms such as FEDAVG are primarily geared towards *smooth unconstrained* settings. In this paper, we study the *Federated Composite Optimization* (FCO) problem, in which the loss function contains a non-smooth regularizer. Such problems arise naturally in FL applications that involve sparsity, low-rank, monotonicity, or more general constraints. We first show that straightforward extensions of primal algorithms such as FEDAVG are not well-suited for FCO since they suffer from the "curse of primal averaging," resulting in poor convergence. As a solution, we propose a new primal-dual algorithm, *Federated Dual Averaging* (FEDDUALAVG), which by employing a novel server dual averaging procedure circumvents the curse of primal averaging. Our theoretical analysis and empirical experiments demonstrate that FEDDUALAVG outperforms the other baselines.

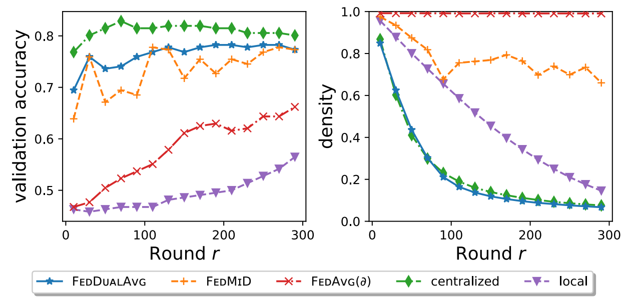

Figure 1. **Results on sparse (ℓ_1-regularized) logistic regression for a federated fMRI dataset based on (Haxby, 2001).** `centralized` corresponds to training on the centralized dataset gathered from **all** the training clients. `local` corresponds to training on the local data from only **one** training client without communication. FEDAVG (∂) corresponds to running FEDAVG algorithms with subgradient in lieu of SGD to handle the non-smooth ℓ_1-regularizer. FEDMID is another straightforward extension of FEDAVG running local proximal gradient method (see Section 3.1 for details). We show that using our proposed algorithm FEDDUALAVG, one can 1) achieve performance comparable to the `centralized` baseline without the need to gather client data, and 2) significantly outperforms the `local` baseline on the isolated data and the FEDAVG baseline. See Section 5.3 for details.

1. Introduction

Federated Learning (FL, Konečný et al. 2015; McMahan et al. 2017) is a novel distributed learning paradigm in which a large number of clients collaboratively train a shared model without disclosing their private local data. The two most distinct features of FL, when compared to classic distributed learning settings, are (1) heterogeneity in data amongst the clients and (2) very high cost to communicate with a client. Due to these aspects, classic distributed optimization algorithms have been rendered ineffective in FL settings (Kairouz et al., 2019). Several algorithms specifically catered towards FL settings have been proposed to address these issues. The most prominent amongst them is Federated Averaging (FEDAVG) algorithm, which by employing local SGD updates, significantly reduces the communication overhead under moderate client heterogeneity. Several follow-up works have focused on improving the FEDAVG in various ways (e.g., Li et al. 2020a; Karimireddy et al. 2020; Reddi et al. 2020; Yuan & Ma 2020).

Existing FL research primarily focuses on the *unconstrained smooth* objectives; however, many FL applications involve non-smooth objectives. Such problems arise naturally in the context of regularization (e.g., sparsity, low-rank, monotonicity, or additional constraints on the model). For instance, consider the problem of cross-silo biomedical FL, where medical organizations collaboratively aim to learn a global model on their patients' data without sharing. In such applications, sparsity constraints are of paramount importance due to the nature of the problem as it involves only a few data samples (e.g., patients) but with very high dimensions (e.g., fMRI scans). For the purpose of illustration, in Fig. 1, we present results on a federated sparse (ℓ_1-regularized) logistic regression task for an fMRI dataset

[1]Stanford University [2]Based on work performed at Google Research [3]Google Research. Correspondence to: Honglin Yuan <yuanhl@stanford.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

(Haxby, 2001). As shown, using a federated approach that can handle non-smooth objectives enables us to find a highly accurate sparse solution without sharing client data.

In this paper, we propose to study the *Federated Composite Optimization* (FCO) problem. As in standard FL, the losses are distributed to M clients. In addition, we assume all the clients share the same, possibly non-smooth, non-finite regularizer ψ. Formally, (FCO) is of the following form

$$\min_{w \in \mathbb{R}^d} \Phi(w) := F(w) + \psi(w) := \frac{1}{M} \sum_{m=1}^{M} F_m(w) + \psi(w), \tag{FCO}$$

where $F_m(w) := \mathbb{E}_{\xi^m \sim \mathcal{D}_m} f(w; \xi^m)$ is the loss at the m-th client, assuming $\mathcal{D}_m$ is its local data distribution. We assume that each client m can access $\nabla f(w; \xi^m)$ by drawing independent samples ξ^m from its local distribution $\mathcal{D}_m$. Common examples of $\psi(w)$ include ℓ_1-regularizer or more broadly ℓ_p-regularizer, nuclear-norm regularizer (for matrix variable), total variation (semi-)norm, etc. The (FCO) reduces to the standard federated optimization problem if $\psi \equiv 0$. The (FCO) also covers the constrained federated optimization if one takes ψ to be the following constraint characteristics $\chi_{\mathcal{C}}(w) := 0$ if $w \in \mathcal{C}$ or $+\infty$ otherwise.

Standard FL algorithms such as FEDAVG (see Algorithm 1) and its variants (e.g., Li et al. 2020a; Karimireddy et al. 2020) are primarily tailored to *smooth unconstrained* settings, and are therefore, not well-suited for FCO. The most straightforward extension of FEDAVG towards (FCO) is to apply local subgradient method (Shor, 1985) in lieu of SGD. This approach is largely ineffective due to the intrinsic slow convergence of subgradient approach (Boyd et al., 2003), which is also demonstrated in Fig. 1 (marked FEDAVG (∂)).

A more natural extension of FEDAVG is to replace the local SGD with proximal SGD (Parikh & Boyd 2014, a.k.a. projected SGD for constrained problems), or more generally, mirror descent (Duchi et al., 2010). We refer to this algorithm as *Federated Mirror Descent* (FEDMID, see Algorithm 2). The most noticeable drawback of a primal-averaging method like FEDMID is the "curse of primal averaging," where the desired regularization of FCO may be rendered completely ineffective due to the server averaging step typically used in FL. For instance, consider a ℓ_1-regularized logistic regression setting. Although each client is able to obtain a sparse solution, simply averaging the client states will inevitably yield a dense solution. See Fig. 2 for an illustrative example.

To overcome this challenge, we propose a novel primal-dual algorithm named *Federated Dual Averaging* (FEDDUALAVG, see Algorithm 3). Unlike FEDMID (or its precursor FEDAVG), the server averaging step of FEDDUALAVG operates in the dual space instead of the primal. Locally, each client runs dual averaging algorithm (Nesterov, 2009) by tracking of a pair of primal and dual states.

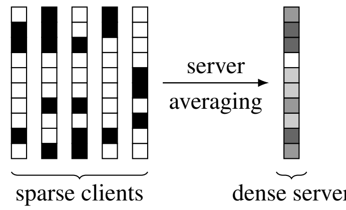

Figure 2. **Illustration of "curse of primal averaging".** While each client of FEDMID can locate a sparse solution, simply averaging them will yield a much denser solution on the server side.

During communication, the dual states are averaged across the clients.

Thus, FEDDUALAVG employs a novel double averaging procedure — averaging of dual states across clients (as in FEDAVG), and the averaging of gradients in dual space (as in the sequential dual averaging). Since both levels of averaging operate in the dual space, we can show that FEDDUALAVG provably overcomes the curse of primal averaging. Specifically, we prove that FEDDUALAVG can attain significantly lower communication complexity when deployed with a large client learning rate.

Contributions. In light of the above discussion, let us summarize our key contributions below:

- We propose a generalized federated learning problem, namely *Federated Composite Optimization* (FCO), with non-smooth regularizers and constraints.
- We first propose a natural extension of FEDAVG, namely *Federated Mirror Descent* (FEDMID). We show that FEDMID can attain the mini-batch rate in the small client learning rate regime (Section 4.1). We argue that FEDMID may suffer from the effect of "curse of primal averaging," which results in poor convergence, especially in the large client learning rate regime (Section 3.2).
- We propose a novel primal-dual algorithm named *Federated Dual Averaging* (FEDDUALAVG), which provably overcomes the curse of primal averaging (Section 3.3). Under certain realistic conditions, we show that by virtue of "double averaging" property, FEDDUALAVG can have significantly lower communication complexity (Section 4.2).
- We demonstrate the empirical performance of FEDMID and FEDDUALAVG on various tasks, including ℓ_1-regularization, nuclear-norm regularization, and various constraints in FL (Section 5).

Notations. We use $[n]$ to denote the set $\{1, \ldots, n\}$. We use $\langle \cdot, \cdot \rangle$ to denote the inner product, $\| \cdot \|$ to denote an arbitrary norm, and $\| \cdot \|_*$ to denote its dual norm, unless otherwise specified. We use $\|\cdot\|_2$ to denote the ℓ_2 norm of a vector or the operator norm of a matrix, and $\|\cdot\|_A$ to denote the vector norm induced by positive definite matrix A, namely $\|w\|_A := \sqrt{\langle w, Aw \rangle}$. For any convex function $g(w)$, we use $g^*(z)$ to denote its convex conjugate $g^*(z) := \sup_{w\in\mathbb{R}^d}\{\langle z, w\rangle - g(w)\}$. We use $w^\star$ to denote the optimum of the problem (FCO). We use $\mathcal{O}, \Theta$ to hide multiplicative absolute constants only and $x \lesssim y$ to denote $x = \mathcal{O}(y)$.

1.1. Related Work

In this subsection, we briefly discuss the main related work. We provide a more detailed literature review in Appendix A, including the relation to classic composite optimization and distributed consensus optimization literature.

The first analysis of general FEDAVG was established by Stich (2019) for the homogeneous client dataset. This result was improved by Haddadpour et al. (2019b); Khaled et al. (2020); Woodworth et al. (2020b); Yuan & Ma (2020) via tighter analysis and accelerated algorithms. For heterogeneous clients, numerous recent papers (Haddadpour et al., 2019b; Khaled et al., 2020; Li et al., 2020b; Koloskova et al., 2020; Woodworth et al., 2020a) studied the convergence of FEDAVG under various notions of heterogeneity measure. FEDAVG has also been studied for non-convex objectives (Zhou & Cong, 2018; Haddadpour et al., 2019a; Wang & Joshi, 2018; Yu & Jin, 2019; Yu et al., 2019a;b). Other variants of FEDAVG have been proposed to overcome heterogeneity challenges (e.g., Mohri et al. 2019; Liang et al. 2019; Li et al. 2020a; Wang et al. 2020; Karimireddy et al. 2020; Pathak & Wainwright 2020; Fallah et al. 2020; Hanzely et al. 2020; T. Dinh et al. 2020; Lin et al. 2020; He et al. 2020; Bistritz et al. 2020; Zhang et al. 2020). We refer readers to (Kairouz et al., 2019) for a comprehensive survey of recent advances in FL.

We note that none of the aforementioned works can handle non-smooth problems such as (FCO). Furthermore, the contributions of this work can potentially be integrated with other emerging techniques in FL (e.g., acceleration, adaptivity, variance reduction) to overcome challenges in FL such as communication efficiency and client heterogeneity.

2. Preliminaries

In this section, we review the necessary background for composite optimization and federated learning. A detailed technical exposition of these topics is relegated to Appendix C.

2.1. Composite Optimization

Composite optimization covers a variety of statistical inference, machine learning, signal processing problems. Standard (non-distributed) composite optimization is defined as

$$\min_{w\in\mathbb{R}^d} \quad \mathbb{E}_{\xi\sim\mathcal{D}}\, f(w;\xi) + \psi(w), \tag{CO}$$

where ψ is a non-smooth, possibly non-finite regularizer.

Proximal Gradient Method. A natural extension of SGD for (CO) is the following *proximal gradient method* (PGM):

$$\begin{aligned} w_{t+1} &\leftarrow \mathbf{prox}_{\eta\psi}\left(w_t - \eta \nabla f(w_t;\xi_t)\right) \\ &= \arg\min_w \left(\eta \langle \nabla f(w_t;\xi_t), w\rangle + \frac{1}{2}\|w - w_t\|_2^2 + \eta\psi(w) \right). \end{aligned} \tag{2.1}$$

The sub-problem Eq. (2.1) can be motivated by optimizing a quadratic upper bound of f together with the original ψ. This problem can often be efficiently solved by virtue of the special structure of ψ. For instance, one can verify that PGM reduces to projected gradient descent if ψ is a constraint characteristic $\chi_{\mathcal{C}}$, soft thresholding if $\psi(w) = \lambda\|w\|_1$, or weight decay if $\psi(w) := \lambda\|w\|_2^2$.

Mirror Descent / Bregman-PGM. PGM can be generalized to the Bregman-PGM if one replaces the Euclidean proximity term by the general Bregman divergence, namely

$$w_{t+1} \leftarrow \arg\min_w \left(\eta \left\langle \nabla f(w_t;\xi_t), w\right\rangle + \eta\psi(w) + D_h(w, w_t)\right), \tag{2.2}$$

where h is a strongly convex distance-generating function, D_h is the Bregman divergence which reduces to Euclidean distance if one takes $h(w) = \frac{1}{2}\|w\|_2^2$. We will still refer to this step as a proximal step for ease of reference. This general formulation (2.2) enables an equivalent primal-dual interpretation:

$$w_{t+1} \leftarrow \nabla(h + \eta\psi)^*(\nabla h(w_t) - \nabla f(w_t;\xi_t)). \tag{2.3}$$

A common interpretation of (2.3) is to decompose it into the following three sub-steps (Nemirovski & Yudin, 1983):

(a) Apply ∇h to carry w_t to a dual state (denoted as z_t).

(b) Update z_t to y_{t+1} with the gradient queried at w_t.

(c) Map y_{t+1} back to primal via $\nabla(h + \eta\psi)^*$.

This formulation is known as the *composite objective mirror descent* (COMID, Duchi et al. 2010), or simply *mirror descent* in the literature (Flammarion & Bach, 2017).

Dual Averaging. An alternative approach for (CO) is the following *dual averaging* algorithm (Nesterov, 2009):

$$z_{t+1} \leftarrow z_t - \eta \nabla f\left(\nabla(h + \eta t\psi)^*(z_t);\xi_t\right). \tag{2.4}$$

Similarly, we can decompose (2.4) into two sub-steps:

(a) Apply $\nabla(h + \eta t \psi)^*$ to map dual state z_t to primal w_t. Note that this sub-step can be reformulated into

$$w_t = \arg\min_w \left(\langle -z_t, w\rangle + \eta t\psi(w) + h(w)\right),$$

which allows for efficient computation for many ψ.

(b) Update z_t to z_{t+1} with the gradient queried at w_t.

Dual averaging is also known as the *"lazy" mirror descent* algorithm (Bubeck, 2015) since it skips the forward mapping (∇h) step. Theoretically, mirror descent and dual averaging often share the similar convergence rates for sequential (CO) (e.g., for smooth convex f, c.f. Flammarion & Bach 2017).

Remark. *There are other algorithms that are popular for certain types of* (CO) *problems. For example,* Frank-Wolfe *method (Frank & Wolfe, 1956; Jaggi, 2013) solves constrained optimization with a linear optimization oracle. Smoothing method (Nesterov, 2005) can also handle non-smoothness in objectives, but is in general less efficient than specialized CO algorithms such as dual averaging (c.f., Nesterov 2018). In this work, we mostly focus on Mirror Descent and Dual Averaging algorithms since they only employ simple proximal oracles such as projection and soft-thresholding. We refer readers to Appendix A.2 for additional related work in composite optimization.*

2.2. Federated Averaging

Federated Averaging (FEDAVG, McMahan et al. 2017) is the *de facto* standard algorithm for Federated Learning with unconstrained smooth objectives (namely $\psi = 0$ for (FCO)). In this work, we follow the exposition of (Reddi et al., 2020) which splits the client learning rate and server learning rate, offering more flexibility (see Algorithm 1).

FEDAVG involves a series of *rounds* in which each round consists of a client update phase and server update phase. We denote the total number of rounds as R. At the beginning of each round r, a subset of clients $\mathcal{S}_r$ are sampled from the client pools of size M. The server state is then broadcast to the sampled client as the client initialization. During the client update phase (highlighted in blue shade), each sampled client runs local SGD for K steps with client learning rate η_c with their own data. We use $w_{r,k}^m$ to denote the m-th client state at the k-th local step of the r-th round. During the server update phase, the server averages the updates of the sampled clients and treats it as a pseudo-anti-gradient Δ_r (Line 9). The server then takes a server update step to update its server state with server learning rate η_s and the pseudo-anti-gradient Δ_r (Line 10).

Algorithm 1 Federated Averaging (FEDAVG)

1: **procedure** FEDAVG (w_0, η_c, η_s)
2: **for** $r = 0, \ldots, R-1$ **do**
3: sample a subset of clients $\mathcal{S}_r \subseteq [M]$
4: **on client for** $m \in \mathcal{S}_r$ **in parallel do**
5: $w_{r,0}^m \leftarrow w_r$ ▷ broadcast client initialization
6: **for** $k = 0, \ldots, K-1$ **do**
7: $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$ ▷ query gradient
8: $w_{r,k+1}^m \leftarrow w_{r,k}^m - \eta_c \cdot g_{r,k}^m$ ▷ client update
9: $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
10: $w_{r+1} \leftarrow w_r + \eta_s \cdot \Delta_r$ ▷ server update

Algorithm 2 Federated Mirror Descent (FEDMID)

1: **procedure** FEDMID (w_0, η_c, η_s)
2: **for** $r = 0, \ldots, R-1$ **do**
3: sample a subset of clients $\mathcal{S}_r \subseteq [M]$
4: **on client for** $m \in \mathcal{S}_r$ **in parallel do**
5: $w_{r,0}^m \leftarrow w_r$ ▷ broadcast *primal* initialization
6: **for** $k = 0, \ldots, K-1$ **do**
7: $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$ ▷ query gradient
8: $w_{r,k+1}^m \leftarrow \nabla (h + \eta_c \psi)^* (\nabla h(w_{r,k}^m) - \eta_c \cdot g_{r,k}^m)$
9: $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (w_{r,K}^m - w_{r,0}^m)$
10: $w_{r+1} \leftarrow \nabla (h + \eta_s \eta_c K \psi)^* (\nabla h(w_r) + \eta_s \cdot \Delta_r)$

3.1. Federated Mirror Descent (FEDMID)

A more natural extension of FEDAVG towards (FCO) is to replace the local SGD steps in FEDAVG with local proximal gradient (mirror descent) steps (2.3). The resulting algorithm, which we refer to as *Federated Mirror Descent* (FEDMID)[1], is outlined in Algorithm 2.

Specifically, we make two changes compared to FEDAVG:

- The client local SGD steps in FEDAVG are replaced with proximal gradient steps (Line 8).
- The server update step is replaced with another proximal step (Line 10).

As a sanity check, for constrained (FCO) with $\psi = \chi_{\mathcal{C}}$, if one takes server learning rate $\eta_s = 1$ and Euclidean distance $h(w) = \frac{1}{2}\|w\|_2^2$, FEDMID will simply reduce to the following parallel projected SGD with periodic averaging:

(a) Each sampled client runs K steps of projected SGD following $w_{r,k+1}^m \leftarrow \mathbf{Proj}_{\mathcal{C}}(w_{r,k}^m - \eta_c g_{r,k}^m)$.

3. Proposed Algorithms for FCO

In this section, we explore the possible solutions to approach (FCO). As mentioned earlier, existing FL algorithms such as FEDAVG and its variants do not solve (FCO). Although it is possible to apply FEDAVG to non-smooth settings by using subgradient in place of the gradient, such an approach is usually ineffective owing to the intrinsic slow convergence of subgradient methods (Boyd et al., 2003).

[1] Despite sharing the same term "prox", FEDMID is fundamentally different from FEDPROX (Li et al., 2020a). The proximal step in FEDPROX was to regularize the client drift caused by heterogeneity, whereas the proximal step in this work is to overcome the non-smoothness of ψ. The problems approached by the two methods are also different – FEDPROX still solves an unconstrained smooth problem, whereas ours concerns with approaches (FCO).

(b) After K local steps, the server simply average the client states following $w_{r+1} \leftarrow \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} w_{r,K}^m$.

3.2. Limitation of FEDMID: Curse of Primal Averaging

Despite its simplicity, FEDMID exhibits a major limitation, which we refer to as "curse of primal averaging": the server averaging step in FEDMID may severely impede the optimization progress. To understand this phenomenon, let us consider the following two illustrative examples:

- **Constrained problem**: Suppose the optimum of the aforementioned constrained problem resides on a non-flat boundary $\mathcal{C}$. Even when each client is able to obtain a local solution *on* the boundary, the average of them will almost surely be *off* the boundary (and hence away from the optimum) due to the curvature.
- **Federated ℓ_1-regularized logistic regression problem**: Suppose each client obtains a local *sparse* solution, simply averaging them across clients will invariably yield a non-sparse solution.

As we will see theoretically (Section 4) and empirically (Section 5), the "curse of primal averaging" indeed hampers the performance of FEDMID.

3.3. Federated Dual Averaging (FEDDUALAVG)

Before we look into the solution of the curse of primal averaging, let us briefly investigate the cause of this effect. Recall that in standard smooth FL settings, server averaging step is helpful because it implicitly pools the stochastic gradients and thereby reduces the variance (Stich, 2019). In FEDMID, however, the server averaging operates on the post-proximal **primal** states, but the gradient is updated in the **dual** space (recall the primal-dual interpretation of mirror descent in Section 2.1). This primal/dual mismatch creates an obstacle for primal averaging to benefit from the pooling of stochastic gradients in dual space. This thought experiment suggests the importance of aligning the gradient update and server averaging.

Building upon this intuition, we propose a novel primal-dual algorithm, named *Federated Dual Averaging* (FEDDUALAVG, Algorithm 3), which provably addresses the curse of primal averaging. The major novelty of FEDDUALAVG, in comparison with FEDMID or its precursor FEDAVG, is to operate the server averaging in the dual space instead of the primal. This facilitates the server to aggregate the gradient information since the gradients are also accumulated in the dual space.

Formally, each client maintains a pair of primal and dual states $(w_{r,k}^m, z_{r,k}^m)$. At the beginning of each client update round, the client dual state is initialized with the server dual state. During the client update stage, each client runs dual averaging steps following (2.4) to update its primal and dual state (highlighted in blue shade). The coefficient of ψ, namely $\tilde{\eta}_{r,k}$, is to balance the contribution from F and ψ. At the end of each client update phase, the *dual updates* (instead of primal updates) are returned to the server. The server then averages the dual updates of the sampled clients and updates the server dual state. We observe that the averaging in FEDDUALAVG is two-fold: (1) averaging of gradients in dual space within a client and (2) averaging of dual states across clients at the server. As we shall see shortly in our theoretical analysis, this novel "double" averaging of FEDDUALAVG in the non-smooth case enables lower communication complexity and faster convergence of FEDDUALAVG under realistic assumptions.

Algorithm 3 Federated Dual Averaging (FEDDUALAVG)

procedure FEDDUALAVG (w_0, η_c, η_s)
  $z_0 \leftarrow \nabla h(w_0)$ ▷ server dual initialization
  for $r = 0, \ldots, R-1$ **do**
    sample a subset of clients $\mathcal{S}_r \subseteq [M]$
    on client for $m \in \mathcal{S}_r$ **in parallel do**
      $z_{r,0}^m \leftarrow z_r$ ▷ broadcast *dual* initialization
      for $k = 0, \ldots, K-1$ **do**
        $\tilde{\eta}_{r,k} \leftarrow \eta_s \eta_c r K + \eta_c k$
        $w_{r,k}^m \leftarrow \nabla (h + \tilde{\eta}_{r,k}\psi)^*(z_{r,k}^m)$ ▷ retrieve primal
        $g_{r,k}^m \leftarrow \nabla f(w_{r,k}^m; \xi_{r,k}^m)$ ▷ query gradient
        $z_{r,k+1}^m \leftarrow z_{r,k}^m - \eta_c g_{r,k}^m$ ▷ client *dual* update
    $\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{m \in \mathcal{S}_r} (z_{r,K}^m - z_{r,0}^m)$
    $z_{r+1} \leftarrow z_r + \eta_s \Delta_r$ ▷ server *dual* update
    $w_{r+1} \leftarrow \nabla (h + \eta_s \eta_c (r+1) K \psi)^*(z_{r+1})$
      ▷ (optional) retrieve server primal state

4. Theoretical Results

In this section, we demonstrate the theoretical results of FEDMID and FEDDUALAVG. We assume the following assumptions throughout the paper. The convex analysis definitions in Assumption 1 are reviewed in Appendix C.

Assumption 1. *Let $\|\cdot\|$ be a norm and $\|\cdot\|_*$ be its dual.*

(a) $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function with closed $\mathbf{dom}\,\psi$. Assume that $\Phi(w) = F(w) + \psi(w)$ attains a finite optimum at $w^\star \in \mathbf{dom}\,\psi$.

(b) $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a Legendre function that is 1-strongly-convex w.r.t. $\|\cdot\|$. Assume $\mathbf{dom}\,h \supset \mathbf{dom}\,\psi$.

(c) $f(\cdot, \xi) : \mathbb{R}^d \to \mathbb{R}$ is a closed convex function that is differentiable on $\mathbf{dom}\,\psi$ for any fixed ξ. In addition, $f(\cdot, \xi)$ is L-smooth w.r.t. $\|\cdot\|$ on $\mathbf{dom}\,\psi$, namely for

any $u, w \in \mathbf{dom}\,\psi$*,*

$$f(u;\xi) \le f(w;\xi) + \langle \nabla f(w;\xi), u-w\rangle + \frac{1}{2}L\|u-w\|^2.$$

(d) ∇f *has* σ^2*-bounded variance over* $\mathcal{D}_m$ *under* $\|\cdot\|_*$ *within* $\mathbf{dom}\,\psi$*, namely for any* $w \in \mathbf{dom}\,\psi$*,*

$$\mathbb{E}_{\xi\sim\mathcal{D}_m} \|\nabla f(w,\xi) - \nabla F_m(w)\|_*^2 \le \sigma^2, \text{ for any } m \in [M]$$

(e) Assume that all the M *clients participate in the client updates for every round, namely* $\mathcal{S}_r = [M]$*.*

Assumption 1(a) & (b) are fairly standard for composite optimization analysis (c.f. Flammarion & Bach 2017). Assumption 1(c) & (d) are standard assumptions in stochastic federated optimization literature (Khaled et al., 2020; Woodworth et al., 2020b). (e) is assumed to simplify the exposition of the theoretical results. All results presented can be easily generalized to the partial participation case.

Remark. *This work focuses on convex settings because the non-convex composite optimization (either* F *or* ψ *non-convex) is noticeably challenging and under-developed* ***even for non-distributed settings****. This is in sharp contrast to non-convex smooth optimization for which simple algorithms such as SGD can readily work. Existing literature on non-convex CO (e.g., Attouch et al. 2013; Chouzenoux et al. 2014; Li & Pong 2015; Bredies et al. 2015) typically relies on non-trivial additional assumptions (such as K-Ł conditions) and sophisticated algorithms. Hence, it is beyond the scope of this work to study non-convex FCO.* [2]

4.1. FEDMID and FEDDUALAVG: Small Client Learning Rate Regime

We first show that both FEDMID and FEDDUALAVG are (asymptotically) at least as good as stochastic mini-batch algorithms with R iterations and batch-size MK when client learning rate η_c is sufficiently small.

Theorem 4.1 (Simplified from Theorem F.1)**.** *Assuming Assumption 1, then for sufficiently small client learning rate* η_c*, and server learning rate* $\eta_s = \Theta(\min\{\frac{1}{\eta_c KL}, \frac{B^{\frac{1}{2}}M^{\frac{1}{2}}}{\eta_c K^{\frac{1}{2}} R^{\frac{1}{2}}\sigma}\})$*, both* FEDDUALAVG *and* FEDMID *can output* $\hat{w}$ *such that*

$$\mathbb{E}\left[\Phi(\hat{w})\right] - \Phi(w^\star) \lesssim \frac{LB}{R} + \frac{\sigma B^{\frac{1}{2}}}{\sqrt{MKR}}, \tag{4.1}$$

where $B := D_h(w^\star, w_0)$*.*

The intuition is that when η_c is small, the client update will not drift too far away from its initialization of the round. Due to space constraints, the proof is relegated to Appendix F.

[2] However, we conjecture that for simple non-convex settings (e.g., optimize non-convex f on a convex set, as tested in Appendix B.5), it is possible to show the convergence and obtain similar advantageous results for FEDDUALAVG.

4.2. FEDDUALAVG with a Larger Client Learning Rate: Usefulness of Local Step

In this subsection, we show that FEDDUALAVG may attain stronger results with a larger client learning rate. In addition to possible faster convergence, Theorems 4.2 and 4.3 also indicate that FEDDUALAVG allows for much broader searching scope of efficient learning rates configurations, which is of key importance for practical purpose.

Bounded Gradient. We first consider the setting with bounded gradient. Unlike unconstrained, the gradient bound may be particularly useful when the constraint is finite.

Theorem 4.2 (Simplified from Theorem D.1)**.** *Assuming Assumption 1 and* $\sup_{w\in\mathbf{dom}\,\psi}\|\nabla f(w,\xi)\|_* \le G$*, then for* FEDDUALAVG *with* $\eta_s = 1$ *and* $\eta_c \le \frac{1}{4L}$*, considering*

$$\hat{w} := \frac{1}{KR}\sum_{r=0}^{R-1}\sum_{k=1}^{K}\left[\nabla\left(h + \tilde{\eta}_{r,k}\psi\right)^*\left(\frac{1}{M}\sum_{m=1}^{M} z_{r,k}^m\right)\right], \tag{4.2}$$

the following inequality holds

$$\mathbb{E}\left[\Phi\left(\hat{w}\right)\right] - \Phi(w^\star) \lesssim \frac{B}{\eta_c KR} + \frac{\eta_c\sigma^2}{M} + \eta_c^2 LK^2G^2,$$

where $B := D_h(w^\star, w_0)$*. Moreover, there exists* η_c *such that*

$$\mathbb{E}\left[\Phi(\hat{w})\right] - \Phi(w^\star) \lesssim \frac{LB}{KR} + \frac{\sigma B^{\frac{1}{2}}}{\sqrt{MKR}} + \frac{L^{\frac{1}{3}}B^{\frac{2}{3}}G^{\frac{2}{3}}}{R^{\frac{2}{3}}}. \tag{4.3}$$

We refer the reader to Appendix D for complete proof details of Theorem 4.2.

Remark. *The result in Theorem 4.2 not only matches the rate by Stich (2019) for smooth, unconstrained* FEDAVG *but also allows for a general non-smooth composite* ψ*, general Bregman divergence induced by* h*, and arbitrary norm* $\|\cdot\|$*. Compared with the small learning rate result Theorem 4.1, the first term in Eq. (4.3) is improved from* $\frac{LB}{R}$ *to* $\frac{LB}{KR}$*, whereas the third term incurs an additional loss regarding infrequent communication. One can verify that the bound Eq. (4.3) is better than Eq. (4.1) if* $R \lesssim \frac{L^2B}{G^2}$*. Therefore, the larger client learning rate may be preferred when the communication is not too infrequent.*

Bounded Heterogeneity. Next, we consider the settings with bounded heterogeneity. For simplicity, we focus on the case when the loss F is quadratic, as shown in Assumption 2. We will discuss other options to relax the quadratic assumption in Section 4.3.

Assumption 2 (Bounded heterogeneity, quadratic)**.**

(a) The heterogeneity of ∇F_m *is bounded, namely*

$$\sup_{w\in\mathbf{dom}\,\psi}\|\nabla F_m(w) - \nabla F(w)\|_* \le \zeta^2, \text{ for any } m \in [M]$$

(b) $F(w) := \frac{1}{2}w^\top Q w + c^\top w$ *for some* $Q \succ 0$.

(c) Assume Assumption 1 is satisfied in which the norm $\|\cdot\|$ *is taken to be the* $\frac{Q}{\|Q\|_2}$*-norm, namely* $\|w\| = \sqrt{\frac{w^\top Q w}{\|Q\|_2}}$.

Remark. *Assumption 2(a) is a standard assumption to bound the heterogeneity among clients (e.g., Woodworth et al. 2020a). Note that Assumption 2 only assumes the objective F to be quadratic. We do not impose any stronger assumptions on either the composite function ψ or the distance-generating function h. Therefore, this result still applies to a broad class of problems such as* LASSO.

The following results hold under Assumption 2. We sketch the proof in Section 4.3 and defer the details to Appendix E.

Theorem 4.3 (Simplified from Theorem E.1)**.** *Assuming Assumption 2, then for* FEDDUALAVG *with* $\eta_{\mathrm{s}} = 1$ *and* $\eta_{\mathrm{c}} \leq \frac{1}{4L}$*, the following inequality holds*

$$\mathbb{E}[\Phi(\hat{w})] - \Phi(w^\star) \lesssim \frac{B}{\eta_{\mathrm{c}} K R} + \frac{\eta_{\mathrm{c}}\sigma^2}{M} + \eta_{\mathrm{c}}^2 L K \sigma^2 + \eta_{\mathrm{c}}^2 L K^2 \zeta^2,$$

where $\hat{w}$ *is the same as defined in Eq. (4.2), and* $B := D_h(w^\star, w_0)$*. Moreover, there exists* η_{c} *such that*

$$\mathbb{E}\left[\Phi(\hat{w})\right] - \Phi(w^\star) \lesssim \frac{LB}{KR} + \frac{\sigma B^{\frac{1}{2}}}{\sqrt{MKR}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} \sigma^{\frac{2}{3}}}{K^{\frac{1}{3}} R^{\frac{2}{3}}} + \frac{L^{\frac{1}{3}} B^{\frac{2}{3}} \zeta^{\frac{2}{3}}}{R^{\frac{2}{3}}}. \tag{4.4}$$

Remark. *The result in Theorem 4.3 matches the best-known convergence rate for smooth, unconstrained* FEDAVG *(Khaled et al., 2020; Woodworth et al., 2020a), while our results allow for general composite ψ and non-Euclidean distance. Compared with Theorem 4.2, the overhead in Eq. (4.4) involves variance σ and heterogeneity ζ but no longer depends on G. The bound Eq. (4.4) could significantly outperform the previous ones when the variance σ and heterogeneity ζ are relatively mild.*

4.3. Proof Sketch and Discussions

In this subsection, we demonstrate our proof framework by sketching the proof for Theorem 4.3.

Step 1: Convergence of Dual Shadow Sequence. We start by characterizing the convergence of the dual shadow sequence $\overline{z_{r,k}} := \frac{1}{M}\sum_{m=1}^{M} z_{r,k}^m$. The key observation for FEDDUALAVG when $\eta_s = 1$ is the following relation

$$\overline{z_{r,k+1}} = \overline{z_{r,k}} - \eta_{\mathrm{c}} \cdot \frac{1}{M}\sum_{m=1}^{M} \nabla f(w_{r,k}^m; \xi_{r,k}^m). \tag{4.5}$$

This suggests that the shadow sequence $\overline{z_{r,k}}$ almost executes a dual averaging update (2.4), but with some perturbed gradient $\frac{1}{M}\sum_{m=1}^{M} \nabla f(w_{r,k}^m; \xi_{r,k}^m)$. To this end, we extend the perturbed iterate analysis framework (Mania et al., 2017) to the dual space. Theoretically we show the following Lemma 4.4, with proof relegated to Appendix D.2.

Lemma 4.4 (Convergence of dual shadow sequence of FEDDUALAVG, simplified version of Lemma D.2)**.** *Assuming Assumption 1, then for* FEDDUALAVG *with* $\eta_{\mathrm{s}} = 1$ *and* $\eta_{\mathrm{c}} \leq \frac{1}{4L}$*, the following inequality holds*

$$\mathbb{E}\left[\Phi\left(\frac{1}{KR}\sum_{r=0}^{R-1}\sum_{k=1}^{K} \nabla\left(h + \tilde{\eta}_{r,k}\psi\right)^*\left(\overline{z_{r,k}}\right)\right)\right] - \Phi(w^\star)$$
$$\leq \underbrace{\frac{B}{\eta_{\mathrm{c}} K R} + \frac{\eta_{\mathrm{c}}\sigma^2}{M}}_{\substack{\text{Rate if synchronized} \\ \text{every iteration}}} + \underbrace{\frac{L}{MKR}\left[\sum_{r=0}^{R-1}\sum_{k=0}^{K-1}\sum_{m=1}^{M} \mathbb{E}\left\|\overline{z_{r,k}} - z_{r,k}^m\right\|_*^2\right]}_{\text{Discrepancy overhead}}. \tag{4.6}$$

The first two terms correspond to the rate when FEDDUALAVG is synchronized every step. The last term corresponds to the overhead for not synchronizing every step, which we call "discrepancy overhead". Lemma 4.4 can serve as a general interface towards the convergence of FEDDUALAVG as it only assumes the blanket Assumption 1.

Remark. *Note that the relation (4.5) is not satisfied by* FEDMID *due to the incommutability of the proximal operator and the the averaging operator, which thereby breaks Lemma 4.4. Intuitively, this means* FEDMID *fails to pool the gradients properly (up to a high-order error) in the absence of communication.* FEDDUALAVG *overcomes the incommutability issue because all the gradients are accumuluated and averaged in the dual space, whereas the proximal step only operates at the interface from dual to primal. This key difference explains the "curse of primal averaging" from the theoretical perspective.*

Step 2: Bounding Discrepancy Overhead via Stability Analysis. The next step is to bound the discrepancy term introduced in Eq. (4.6). Intuitively, this term characterizes the *stability* of FEDDUALAVG, in the sense that how far away a single client can deviate from the average (in dual space) if there is no synchronization for k steps.

However, unlike the smooth convex unconstrained settings in which the stability of SGD is known to be well-behaved (Hardt et al., 2016), the stability analysis for composite optimization is challenging and absent from the literature. We identify that the main challenge originates from the asymmetry of the Bregman divergence. In this work, we provide a set of simple conditions, namely Assumption 2, such that the stability of FEDDUALAVG is well-behaved.

Lemma 4.5 (Dual stability of FEDDUALAVG under Assumption 2, simplified version of Lemma E.2)**.** *Under the same settings of Theorem 4.3, the following inequality holds* $\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\left\|\overline{z_{r,k}} - z_{r,k}^m\right\|_*^2 \lesssim \eta_{\mathrm{c}}^2 K\sigma^2 + \eta_{\mathrm{c}}^2 K^2\zeta^2.$

Step 3: Deciding η_{c}. The final step is to plug in the bound in step 2 back to step 1, and find appropriate η_{c} to optimize such upper bound. We defer the details to Appendix E.

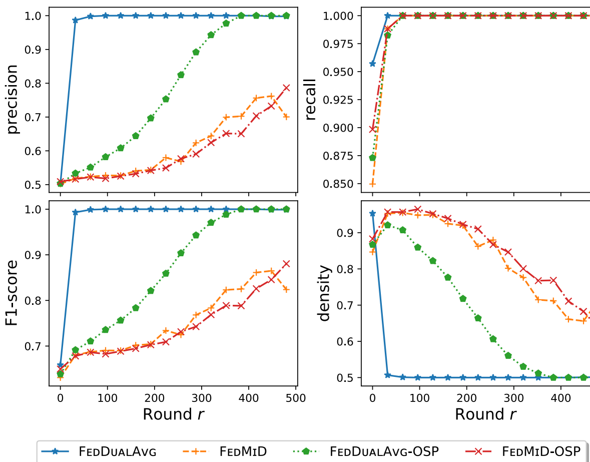

Figure 3. **Sparsity recovery on a synthetic LASSO problem with 50% sparse ground truth.** Observe that FEDDUALAVG not only identifies most of the sparsity pattern but also is fastest. It is also worth noting that the less-principled FEDDUALAVG-OSP is also very competitive. The poor performance of FEDMID can be attributed to the "curse of primal averaging", as the server averaging step "smooths out" the sparsity pattern, which is corroborated empirically by the least sparse solution obtained by FEDMID.

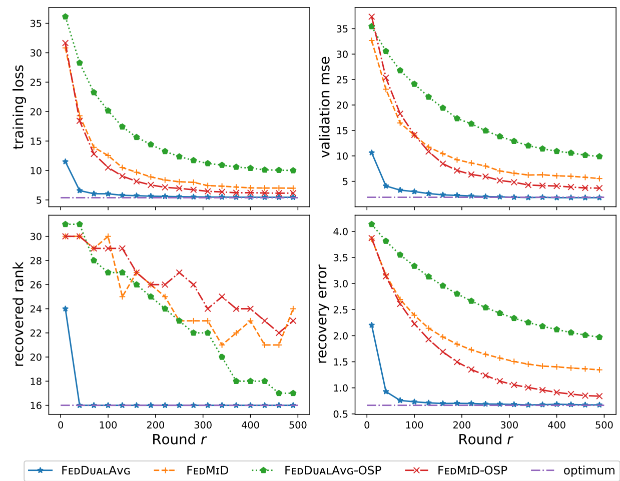

Figure 4. **Low-rank matrix estimation comparison on a synthetic dataset with the ground truth of rank 16.** We observe that FEDDUALAVG finds the solution with exact rank in less than 100 communication rounds. FEDMID and FEDMID-OSP converge slower in loss and rank. The unprincipled FEDDUALAVG-OSP can generate low-rank solutions but is far less accurate.

5. Numerical Experiments

In this section, we validate our theory and demonstrate the efficiency of the algorithms via numerical experiments. We mostly compare FEDDUALAVG with FEDMID since the latter serves a natural baseline. We do not present subgradient-FEDAVG in this section due to its consistent ineffectiveness, as demonstrated in Fig. 1 (marked FEDAVG (∂)). To examine the necessity of client proximal step, we also test two less-principled versions of FEDMID and FEDDUALAVG, in which the proximal steps are only performed on the server-side. We refer to these two versions as FEDMID-OSP and FEDDUALAVG-OSP, where "OSP" stands for "only server proximal," with pseudo-code provided in Appendix B.1. We provide the complete setup details in Appendix B, including but not limited to hyper-parameter tuning, dataset processing and evaluation metrics. The source code is available at https://github.com/hongliny/FCO-ICML21.

5.1. Federated LASSO for Sparse Feature Recovery

In this subsection, we consider the LASSO (ℓ_1-regularized least-squares) problem on a synthetic dataset, motivated by models from biomedical and signal processing literature (e.g., Ryali et al. 2010; Chen et al. 2012). The goal is to recover the sparse signal w from noisy observations (x, y).

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}_m} (x^\top w + b - y)_2^2 + \lambda \|w\|_1.$$

To generate the synthetic dataset, we first fix a sparse ground truth $w_{\text{real}} \in \mathbb{R}^d$ and some bias $b_{\text{real}} \in \mathbb{R}$, and then sample the dataset (x, y) following $y = x^\top w_{\text{real}} + b_{\text{real}} + \varepsilon$ for some noise ε. We let the distribution of (x, y) vary over clients to simulate the heterogeneity. We select λ so that the centralized solver (on gathered data) can successfully recover the sparse pattern. Since the ground truth w_{real} is known, we can assess the quality of the sparse features recovered by comparing it with the ground truth.

We evaluate the performance by recording precision, recall, sparsity density, and F1-score. We tune the client learning rate η_c and server learning rate η_s only to attain the best F1-score. The results are presented in Fig. 3. The best learning rates configuration is $\eta_c = 0.01, \eta_s = 1$ for FEDDUALAVG, and $\eta_c = 0.001, \eta_s = 0.3$ for other algorithms (including FEDMID). This matches our theory that FEDDUALAVG can benefit from larger learning rates. We defer the rest of the setup details and further experiments to Appendix B.2.

5.2. Federated Low-Rank Matrix Estimation via Nuclear-Norm Regularization

In this subsection, we consider a low-rank matrix estimation problem via the nuclear-norm regularization

$$\min_{W,b} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{(X,y)\sim\mathcal{D}_m} \left(\langle X, W\rangle + b - y\right)^2 + \lambda \|W\|_{\text{nuc}},$$

where $\|W\|_{\text{nuc}}$ denotes the matrix nuclear norm. The goal is to recover a low-rank matrix W from noisy observations

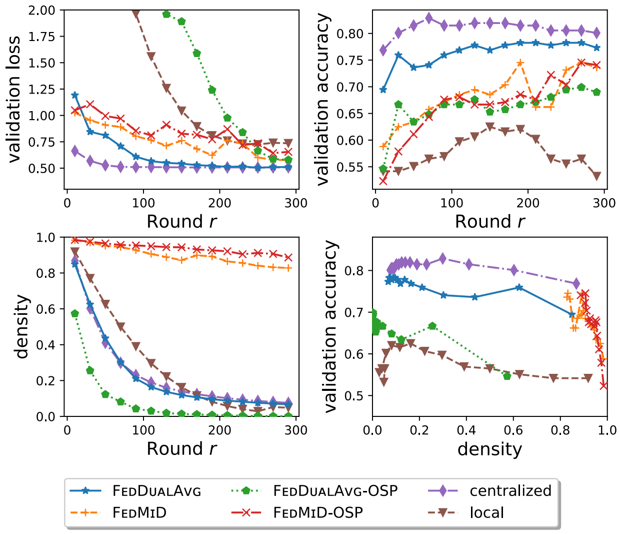

Figure 5. **Results on ℓ_1-regularized logistic regression for fMRI data from (Haxby, 2001).** We observe that FEDDUALAVG yields sparse and accurate solutions that are comparable with the centralized baseline. FEDMID and FEDMID-OSP provides denser solutions that are relatively less accurate. The unprincipled FEDDUALAVG-OSP can provide sparse solutions but far less accurate.

(X, y). This formulation captures a variety of problems such as low-rank matrix completion and recommendation systems (Candès & Recht, 2009). Note that the proximal operator with respect to the nuclear-norm regularizer reduces to singular-value thresholding operation (Cai et al., 2010).

We evaluate the algorithms on a synthetic federated dataset with known low-rank ground truth $W_{\text{real}} \in \mathbb{R}^{d_1 \times d_2}$ and bias $b_{\text{real}} \in \mathbb{R}$, similar to the above LASSO experiments. We focus on four metrics for this task: the training (regularized) loss, the validation mean-squared-error, the recovered rank, and the recovery error in Frobenius norm $\|W_{\text{output}} - W_{\text{real}}\|_{\text{F}}$. We tune the client learning rate η_{c} and server learning rate η_{s} only to attain the best recovery error. We also record the results obtained by the deterministic solver on centralized data, marked as `optimum`. The results are presented in Fig. 4. We provide the rest of the setup details and more experiments in Appendix B.3.

5.3. Sparse Logistic Regression for fMRI Scan

In this subsection, we consider the cross-silo setup of learning a binary classifier on fMRI scans. For this purpose, we use the data collected by Haxby (2001), to understand the pattern of response in the ventral temporal (vt) area of the brain given a visual stimulus. There were six subjects doing image recognition in a block-design experiment over 11 to 12 sessions, with a total of 71 sessions. Each session consists of 18 fMRI scans under the stimuli of a picture of either a house or a face. We use the `nilearn` package (Abraham et al., 2014) to normalize and transform the four-dimensional raw fMRI scan data into an array with 39,912 volumetric pixels (voxels) using the standard mask. We plan to learn a sparse (ℓ_1-regularized) binary logistic regression on the voxels to classify the stimuli given the voxels input. Enforcing sparsity is crucial for this task as it allows domain experts to understand which part of the brain is differentiating between the stimuli. We select five (out of six) subjects as the training set and the last subject as the held-out validation set. We treat each session as a client, with a total of 59 training clients and 12 validation clients, where each client possesses the voxel data of 18 scans. As in the previous experiment, we tune the client learning rate η_{c} and server learning rate η_{s} only. We set the ℓ_1-regularization strength to be 10^{-3}. For each setup, we run the federated algorithms for 300 communication rounds.

We compare the algorithms with two non-federated baselines: 1) `centralized` corresponds to training on the centralized dataset gathered from **all** the training clients; 2) `local` corresponds to training on the local data from only **one** training client without communication. The results are shown in Fig. 5. In Appendix B.4.2, we provide another presentation of this experiment to visualize the progress of federated algorithms and understand the robustness to learning rate configurations. The results suggest FEDDUALAVG not only recovers sparse and accurate solutions, but also behaves most robust to learning-rate configurations. We defer the rest of the setup details to Appendix B.4.

In Appendix B.5, we provide another set of experiments on federated constrained optimization for Federated EMNIST dataset (Caldas et al., 2019).

6. Conclusion

In this paper, we have shown the shortcomings of primal FL algorithms for FCO and proposed a primal-dual method (FEDDUALAVG) to tackle them. Our theoretical and empirical analysis provide strong evidence to support the superior performance of FEDDUALAVG over natural baselines. Potential future directions include control variates and acceleration based methods for FCO, and applying FCO to personalized settings.

Acknowledgement

We would like to thank Zachary Charles, Zheng Xu, Andrew Hard, Ehsan Amid, Amr Ahmed, Aranyak Mehta, Qian Li, Junzi Zhang, Tengyu Ma, and TensorFlow Federated team for helpful discussions at various stages of this work. Honglin Yuan would like to thank the support by the TOTAL Innovation Scholars program. We would like to thank the anonymous reviewers for their suggestions and comments.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014.

Al-Shedivat, M., Gillenwater, J., Xing, E., and Rostamizadeh, A. Federated Learning via Posterior Averaging: A New Perspective and Practical Algorithms. In *International Conference on Learning Representations*, 2021.

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2), 2013.

Bauschke, H. H., Borwein, J. M., et al. Legendre functions and the method of random Bregman projections. *Journal of convex analysis*, 4(1), 1997.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3), 2003.

Bistritz, I., Mann, A., and Bambos, N. Distributed Distillation for On-Device Learning. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020.

Boyd, S., Xiao, L., and Mutapcic, A. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004, 2003.

Bredies, K., Lorenz, D. A., and Reiterer, S. Minimization of Non-smooth, Non-convex Functionals by Iterative Thresholding. *Journal of Optimization Theory and Applications*, 165(1), 2015.

Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3), 1967.

Bubeck, S. *Convex Optimization: Algorithms and Complexity*. 2015.

Cai, J.-F., Candès, E. J., and Shen, Z. A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM Journal on Optimization*, 20(4), 2010.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., and Talwalkar, A. LEAF: A Benchmark for Federated Settings. In *NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

Candès, E. J. and Recht, B. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6), 2009.

Chen, F., Luo, M., Dong, Z., Li, Z., and He, X. Federated Meta-Learning with Fast Convergence and Efficient Communication. *arXiv:1802.07876 [cs]*, 2019.

Chen, X., Lin, Q., and Pena, J. Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012.

Chouzenoux, E., Pesquet, J.-C., and Repetti, A. Variable Metric Forward–Backward Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function. *Journal of Optimization Theory and Applications*, 162(1), 2014.

Daubechies, I., Defrise, M., and De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11), 2004.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(6), 2012.

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive Personalized Federated Learning. *arXiv:2003.13461 [cs, stat]*, 2020.

Diakonikolas, J. and Orecchia, L. The Approximate Duality Gap Technique: A Unified Theory of First-Order Methods. *SIAM Journal on Optimization*, 29(1), 2019.

Duchi, J. and Singer, Y. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(99), 2009.

Duchi, J. C., Shalev-shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. In *COLT 2010*, 2010.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling. *IEEE Transactions on Automatic Control*, 57(3), 2012.

Fallah, A., Mokhtari, A., and Ozdaglar, A. E. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In *Advances in Neural Information Processing Systems 33*, 2020.

Flammarion, N. and Bach, F. Stochastic composite least-squares regression with convergence rate O(1/n). In *Proceedings of the 2017 Conference on Learning Theory*, volume 65. PMLR, 2017.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2), 1956.

Godichon-Baggioni, A. and Saadane, S. On the rates of convergence of parallelized averaged stochastic gradient algorithms. *Statistics*, 54(3), 2020.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019a.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019b.

Hallac, D., Leskovec, J., and Boyd, S. Network Lasso: Clustering and Optimization in Large Graphs. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

Hanzely, F., Hanzely, S., Horváth, S., and Richtárik, P. Lower bounds and optimal algorithms for personalized federated learning. In *Advances in Neural Information Processing Systems 33*, 2020.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated Learning for Mobile Keyboard Prediction. *arXiv:1811.03604 [cs]*, 2018.

Hard, A., Partridge, K., Nguyen, C., Subrahmanya, N., Shah, A., Zhu, P., Lopez-Moreno, I., and Mathews, R. Training keyword spotting models on non-iid data with federated learning. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. ISCA, 2020.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48. PMLR, 2016.

Hartmann, F., Suh, S., Komarzewski, A., Smith, T. D., and Segall, I. Federated Learning for Ranking Browser History Suggestions. *arXiv:1911.11807 [cs, stat]*, 2019.

Haxby, J. V. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, 293(5539), 2001.

He, C., Annavaram, M., and Avestimehr, S. Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge. In *Advances in Neural Information Processing Systems 33*, volume 33, 2020.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Fundamentals of Convex Analysis*. Springer Berlin Heidelberg, 2001.

Ingerman, A. and Ostrowski, K. Introducing TensorFlow Federated, 2019.

Jaggi, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28. PMLR, 2013.

Jain, P., Kakade, S. M., Kidambi, R., Netrapalli, P., and Sidford, A. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of the 31st Conference on Learning Theory*, volume 75. PMLR, 2018.

Jiang, Y., Konečný, J., Rush, K., and Kannan, S. Improving Federated Learning Personalization via Model Agnostic Meta Learning. *arXiv:1909.12488 [cs, stat]*, 2019.

Jung, A. On the Complexity of Sparse Label Propagation. *Frontiers in Applied Mathematics and Statistics*, 4, 2018.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the International Conference on Machine Learning 1 Pre-Proceedings (ICML 2020)*, 2020.

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter Theory for Local SGD on Identical and Heterogeneous Data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108. PMLR, 2020.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. A Unified Theory of Decentralized SGD with Changing Topology and Local Updates. In *Proceedings of the International Conference on Machine Learning 1 Pre-Proceedings (ICML 2020)*, 2020.

Konečný, J., McMahan, B., and Ramage, D. Federated optimization: Distributed optimization beyond the datacenter. In *8th NIPS Workshop on Optimization for Machine Learning*, 2015.

Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10(28), 2009.

Li, G. and Pong, T. K. Global Convergence of Splitting Methods for Nonconvex Composite Optimization. *SIAM Journal on Optimization*, 25(4), 2015.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020*, 2020a.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of FedAvg on non-iid data. In *International Conference on Learning Representations*, 2020b.

Liang, X., Shen, S., Liu, J., Pan, Z., Chen, E., and Cheng, Y. Variance Reduced Local SGD with Lower Communication Complexity. *arXiv:1912.12844 [cs, math, stat]*, 2019.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble Distillation for Robust Model Fusion in Federated Learning. In *Advances in Neural Information Processing Systems 33*, 2020.

Liu, S., Chen, P.-Y., and Hero, A. O. Accelerated Distributed Dual Averaging Over Evolving Networks of Growing Connectivity. *IEEE Transactions on Signal Processing*, 66(7), 2018.

Lu, H., Freund, R. M., and Nesterov, Y. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1), 2018.

Mahdavi, M., Yang, T., Jin, R., Zhu, S., and Yi, J. Stochastic gradient descent with only one projection. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012.

Mania, H., Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., and Jordan, M. I. Perturbed Iterate Analysis for Asynchronous Stochastic Optimization. *SIAM Journal on Optimization*, 27(4), 2017.

Mcdonald, R., Mohri, M., Silberman, N., Walker, D., and Mann, G. S. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in Neural Information Processing Systems 22*. Curran Associates, Inc., 2009.

McMahan, B. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15. PMLR, 2011.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54. PMLR, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

Mokhtari, A. and Ribeiro, A. DSA: Decentralized double stochastic averaging gradient algorithm. *Journal of Machine Learning Research*, 17(61), 2016.

Nedic, A. and Ozdaglar, A. Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE Transactions on Automatic Control*, 54(1), 2009.

Nedich, A. *Convergence Rate of Distributed Averaging Dynamics and Optimization in Networks*, volume 2. 2015.

Nemirovski, A. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1), 2005.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1), 2009.

Nesterov, Y. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1), 2013.

Nesterov, Y. *Lectures on Convex Optimization*. 2018.

Parikh, N. and Boyd, S. P. *Proximal Algorithms*, volume 1. Now Publishers Inc., 2014.

Pathak, R. and Wainwright, M. J. FedSplit: An algorithmic framework for fast federated optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

Rabbat, M. Multi-agent mirror descent for decentralized stochastic optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2015.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive Federated Optimization. *arXiv:2003.00295 [cs, math, stat]*, 2020.

Rockafellar, R. T. *Convex Analysis*. Number 28. Princeton University Press, 1970.

Rosenblatt, J. D. and Nadler, B. On the optimality of averaging in distributed statistical learning. *Information and Inference*, 5(4), 2016.

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. Sparse logistic regression for whole-brain classification of fMRI data. *NeuroImage*, 51(2), 2010.

Shamir, O. and Srebro, N. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2014.

Shi, W., Ling, Q., Wu, G., and Yin, W. EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization. *SIAM Journal on Optimization*, 25(2), 2015a.

Shi, W., Ling, Q., Wu, G., and Yin, W. A Proximal Gradient Algorithm for Decentralized Composite Optimization. *IEEE Transactions on Signal Processing*, 63(22), 2015b.

Shor, N. Z. *Minimization Methods for Non-Differentiable Functions*, volume 3. Springer Berlin Heidelberg, 1985.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.

Stich, S. U. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Sundhar Ram, S., Nedić, A., and Veeravalli, V. V. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization. *Journal of Optimization Theory and Applications*, 147(3), 2010.

T. Dinh, C., Tran, N., and Nguyen, T. D. Personalized Federated Learning with Moreau Envelopes. In *Advances in Neural Information Processing Systems 33*, 2020.

Tong, Q., Liang, G., Zhu, T., and Bi, J. Federated Nonconvex Sparse Learning. *arXiv:2101.00052 [cs]*, 2020.

Tsianos, K. I. and Rabbat, M. G. Distributed dual averaging for convex optimization under communication delays. In *2012 American Control Conference (ACC)*. IEEE, 2012.

Tsianos, K. I., Lawlor, S., and Rabbat, M. G. Push-Sum Distributed Dual Averaging for convex optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012.

Wang, J. and Joshi, G. Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. *arXiv:1808.07576 [cs, stat]*, 2018.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. SlowMo: Improving communication-efficient distributed SGD with slow momentum. In *International Conference on Learning Representations*, 2020.

Woodworth, B., Patel, K. K., and Srebro, N. Minibatch vs Local SGD for Heterogeneous Distributed Learning. In *Advances in Neural Information Processing Systems 33*, 2020a.

Woodworth, B., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., and Srebro, N. Is Local SGD Better than Minibatch SGD? In *Proceedings of the International Conference on Machine Learning 1 Pre-Proceedings (ICML 2020)*, 2020b.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(88), 2010.

Yang, T., Lin, Q., and Zhang, L. A richer theory of convex constrained optimization with reduced projections and improved rates. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70. PMLR, 2017.

Yu, H. and Jin, R. On the computation and communication complexity of parallel SGD with dynamic batch sizes for stochastic non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97. PMLR, 2019a.

Yu, H., Yang, S., and Zhu, S. Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019b.

Yuan, D., Hong, Y., Ho, D. W., and Jiang, G. Optimal distributed stochastic mirror descent for strongly convex optimization. *Automatica*, 90, 2018.

Yuan, D., Hong, Y., Ho, D. W. C., and Xu, S. Distributed Mirror Descent for Online Composite Optimization. *IEEE Transactions on Automatic Control*, 2020.

Yuan, H. and Ma, T. Federated Accelerated Stochastic Gradient Descent. In *Advances in Neural Information Processing Systems 33*, 2020.

Yuan, K., Ling, Q., and Yin, W. On the Convergence of Decentralized Gradient Descent. *SIAM Journal on Optimization*, 26(3), 2016.

Zhang, X., Hong, M., Dhople, S., Yin, W., and Liu, Y. FedPD: A Federated Learning Framework with Optimal Rates and Adaptivity to Non-IID Data. *arXiv:2005.11418 [cs, stat]*, 2020.

Zhou, F. and Cong, G. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*. AAAI Press, 2003.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. J. Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010.